UTILIZING MULTIPLE LINEAR REGRESSION TECHNIQUE FOR INTERENTIAL MEASURE OF CONTINUOUS-BASED PROCESS MONITORING

MUHAMMAD RIDZUAN BIN MAMAT

Thesis submitted in partial fulfilment of the requirements
for the award of the degree of
Bachelor of Chemical Engineering

Faculty of Chemical & Natural Resources Engineering
UNIVERSITI MALAYSIA PAHANG

JULY 2013

ABSTRACT

The present conventional MSPC has several weaknesses in process fault detection and diagnosis. Some researchers in this filed had commented that the MSPC is a powerful tool for data complexity reduction and fault detection in the significant fault appearance data. The current fault detection and diagnosis method via MSPC is limited to significant faults and does not point put the insignificant ones accurately. In the real time, all variable will be used in monitoring. However in this case only a few of them are truly important. By developed modeling based on multiple linear regressions the relationship between these variables can be figure out. Multiple linear regressions (MLR) is a method used to model the linear relationship between a dependent variable and one or more independent variables. Some assumption should be made in order to obtain an accurate data analysis. The assumptions are variables should normally distribute, a linear relationship between the independent and dependent variables must exist and also the variable should be measure without an error. MLR is probably the most widely used in dendroclimatology for developing models to reconstruct climate variables. Besides they also proposed for control charting methods for lumber manufacturing and profile monitoring applied in public health surveillance. The methods to perform this modeling involve two phases which are Phase I: offline modeling and monitoring and Phase II: online monitoring. As a conclusion, the MLR method is success introduced as a significant improvement compared to the conventional method. On top of those objectives, the original goals of SPC are also been considered as well as carried together, such a way that the productivity of multivariate process monitoring is improved.

ABSTRAK

Kaedah konvensional MSPC yang digunakan kini mempunyai beberapa kelemahan di dalam diagnosis dan proses menentukan kesalahan. Sesetengah pengkaji di dalam bidang ini menyatakan MSPC adalah satu kaedah yang terbaik bagi pengurangan kerumitan data dan pengesahan kesalahan. Kaedah terkini dalam menentukan kesalahan dan diagnosis melalui MSPC adalah terhad kepada kesalahan yang lebih penting dan dominan. Dalam kes sebenar, semua pembolehubah akan digunakan di dalam process pemantauan. Walaubagaimanapun hanya segelintir pembolehubah yang benar-benar memainkan peranan dalam menetukan kesalahan. Dengan menghasilkan model berasaskan MLR, hubungan antara pembolehubah dapat dikenalpasti. 'Multiple Liner Regression' (MLR) merupakan satu kaedah yang digunakan untuk melihat perkaitan antara pembolehubah yang bersandar and yang tidak bersandar antara satu sama lain. Beberapa andaian perlu di lakukan dan dikenal pasti bagi meperolehi data analisis yang tepat. Antara andaian yang perlu dilakukan adalah taburan pembolehubah haruslah di serakan secara normal, mesti wujud perkaitan di antara pembolehubah yang bersandar debgan yang tidak bersandar dan pembolehubah seharusnya di ambil tanpa sebarang ralat. MLR banyak dia gunakan di dalam bidang dendroclimatology dengan membina model berasaskan pembolehubah iklim yang telah di bina semula. Selain daripada itu, ia juga dicadangkan untuk gunakan bagi kaedah carta kawalan di dalam sector pembalakan dan pemantauan kesihatan. Bagi menghasilkan medel ini, terdapat dua frasa iaitu Frasa I dan Frasa II. Kesimpulanya, kaedah MLR adalah berjaya di perkenalkan dalam membaikpulih kaeadah penentuan kesalahan berbanding kaeadah konvensional yang digunakan sekarang.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

DISSIM	Dissimilarity Method
FCCU	Fluidize Catalytic Cracker Unit
MLR	Multiple Linear Regression
MSPC	Multivariate Statistical Process Control
NOC	Normal Operating Condition
OLS	Ordinary Least Square
PC	Principle Component
PCA	Principle Component Analysis
SPE	Square Prediction Error

Notation:

$C_{m\times m}$	Variance covariance matrix
R^2	Multiple correlation coefficients
b_0	Regression constant
b_k	Coefficient on the kth predictor
e_i	Error term
‘	Transpose, e.g V'
$*$	Multiplication, e.g A^*B
i	Index object
K	Total number predictor
m	Index of Y-variables (m=1,2…M)
N	Number of object
T	Score matrix
X	Matrix of prediction variables
Y	Matrix of response variables

1 INTRODUCTION

1.1 Motivation and statement of problem

Practically, all variable will be used in monitoring. However in this case only a few of them are truly important. By developed modeling based on multiple linear regressions the relationship between these variables can be figure out.

The present conventional MSPC has several weaknesses in process fault detection and diagnosis. Some researchers in this filed had commented that the MSPC is a powerful tool for data complexity reduction and fault detection in the significant fault appearance data. According to Manabu and his research partner, (2000), the current fault detection and diagnosis method via MSPC is limited to significant faults and does not point put the insignificant ones accurately. Qin (2001), also commented that the contribution chart does not have a control limit, making it difficult to determine what is the root cause of the abnormal operating condition.

As a summary of summary other researchers, the weakness of the conventional MSPC can be briefly concluded into three disadvantages. First of all, the complicated control charts are not 'user-friendly', secondly, the conventional MSPC fault detection tools are easily rise up to noisy-fault –signals and lastly, the conventional fault diagnosis is not ready with a proper control limit, thus it cannot determine the root cause of the fault, especially multiple faults. In order to improve the limitation of MSPC, this research should focus on the alternative, which can solve the disadvantages mentioned above.

1.2 Objectives

The following are the objectives of this research:

i. To develop modeling based on Multiple Linear Regression technique for process monitoring.

ii. To study the relationship between explanatory variables and a response variable based on model developed.

1.3 Scope of this research

There are several major scopes that have been identified in order to achieve the objectives of this research:

i. Generate a set of normal operating condition (NOC) data
ii. Develop conventional MSPC on-line monitoring systems for process fault detection and diagnosis
iii. Develop modified MSPC on-line monitoring systems for process fault detection and diagnosis
iv. Improving the conventional contribution chart for fault diagnosis purpose.

1.4 Rational and Significance

In this research, effort mainly concentrates on breaking through the current limitation and the further application of MSPC on a multivariate continuous chemical process. The main contributions of this research are:

i. Application of MSPC tools on the fault detection and diagnosis.
ii. An Eigenvalue-eigenvector PCA approach had been used for developing Principals Components model.
iii. Applied Multiple Linear Regression technique to performing process monitoring.
iv. Modified Process Fault Detection and Diagnosis, mechanisms are also developed based on the Outline Analysis.

1.5 Main contribution of this work

The following are the contributions of this study which are to develop a model based on the Multiple Linear Regression technique to implement in the process monitoring.

1.6 Organisation of this thesis

The structure of the reminder of the thesis is outlined as follow:

Chapter 2 provides a description of the fundamental Principle Component Analysis (PCA). PCA is one of the most common multivariate analyses applied in the Multivariate Statistical Process Control (MSPC). In particular, PCA can be described by means of either using mathematical representation or graphical representation. Also Under this chapter the principle of Multiple Linear Regression (MLR) technique is discussed how it implements as tools in the process monitoring. This chapter also review about the previous research has been done by the other researcher related to the title including the limitation of the MLR and its applications.

Chapter 3 gives a review the methods approached on how the normal distribute data is transform into MSPC for performing process monitoring. It also detail explained on how MLR involve in this method step by step.

Chapter 4 is devoted to discussed the result obtained based on the charts. The comparison between charts is made in order to figure out the result. The abrupt fault and incipient chart is used to figure out at which variables the fault is occur.

Chapter 5 draws together a summary of the thesis and outlines the future work which might be derived from the model developed in this work.

2 LITERATURE REVIEW

2.1 Overview

In this topic, the principle of Multiple Linear Regression (MLR) will be discussion in which this mathematical technique will be used as tools to analyse a set of data from the real chemical process into graph form.

2.2 Fundamental of Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the most common multivariate analyses applied in the MSPC area (Jackson, 1991, MacGregor and Kourti, 1995; Zhang *et al.*, 1997, Gnanadesikan, 1997). In particular, PCA can be described by means of either using mathematical representation or graphical representation. Firstly, from the mathematical point of view, PCA is a multivariate projection technique, which can transform a set of original variables $[\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_m]$ to a set of new variables$[\mathbf{p_1} \quad \mathbf{p_2} \quad \cdots \quad \mathbf{p_m}]$. Generally, these newly formed variables are called Principal Components, PC for short (Jackson, 1991), whereby they build the individual linear combinations of the original variables which are simplified as follows:

$$\mathbf{P}_{n\times m} = \mathbf{X}_{n\times m}\mathbf{V}_{m\times m} \tag{2.1}$$

Initially, the original matrix, **X** has *m* variables with each variable has *n* number of measurements. The data are arranged in the form of *n* x *m*, where the measurements of a variable are organized in the form of a column vector, which is shown as follows:

$$\mathbf{X}_{n\times m} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_m] = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \tag{2.2}$$

Next, **V** is known as the eigenvector matrix, whereby it gives the weighting function in forming the linear combinations of the original variables. The eigenvector matrix, **V**, contains eigenvectors or also known as loading vectors $\mathbf{v_1}$, $\mathbf{v_2}$, …, $\mathbf{v_m}$. Each eigenvector $\mathbf{v_m}$ is a column vector which contains the arrangement of elements $\mathbf{v}^T_{\boldsymbol{m}} = [v_{1,m}, v_{2,m} \ldots v_{m,m}]^T$, as denoted in the following matrix:

$$\mathbf{V}_{m\times m}, =[\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_m]=\begin{bmatrix} v_{1,1} & v_{1,2} & \dots & v_{1,m} \\ v_{2,1} & v_{2,2} & \dots & v_{2,m} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ v_{m,1} & v_{m,2} & \dots & v_{m,m} \end{bmatrix} \tag{2.3}$$

Lastly, **P** is the principal components scores matrix, in which it contains *n* scores for each of the principal components, as given subsequently as:

$$\mathbf{P} = [\mathbf{p}_1 \quad \dots \quad \mathbf{p}_m] \tag{2.4}$$

$$= \begin{bmatrix} x_{1,1}v_{1,1} + x_{1,2}v_{2,1} + \cdots + x_{1,m}v_{m,1} & \dots & x_{1,1}v_{1,m} + x_{1,2}v_{2,m} + \cdots + x_{1,m}v_{m,m} \\ x_{2,1}v_{1,1} + x_{2,2}v_{2,1} + \cdots + x_{2,m}v_{m,1} & \dots & x_{2,1}v_{1,m} + x_{2,2}v_{2,m} + \cdots + x_{2,m}v_{m,m} \\ \vdots & \vdots & \vdots \\ x_{n,1}v_{1,1} + x_{n,2}v_{2,1} + \cdots + x_{n,m}v_{m,1} & \dots & x_{n,1}v_{1,m} + x_{n,2}v_{2,m} + \cdots + x_{n,m}v_{m,m} \end{bmatrix} \tag{2.5}$$

Thus, as far as multivariate calculations are concerned, **P**, will play the role of quality variables rather than individual variables of **X**, in the MSPC method. The original data, **X** are also could be predictable backed from the calculated PC element, by which, the procedures are:

$$\mathbf{X}_{n\times m}\mathbf{V}_{m\times m} = \mathbf{P}_{n\times m} \tag{2.6}$$

$$\mathbf{X}_{n\times m}\mathbf{V}_{m\times m}\mathbf{V}^T_{mxm} = \mathbf{P}_{n\times m}\mathbf{V}^T_{mxm} \tag{2.7}$$

$$\mathbf{X}_{n\times m}\mathbf{I}_{m\times m} = \mathbf{P}_{n\times m}\mathbf{V}^T_{mxm} \tag{2.8}$$

$$\mathbf{X}_{n\times m} = [\mathbf{p_1} \quad \mathbf{p_2} \quad \dots \quad \mathbf{p}_m]\begin{bmatrix} \mathbf{v_1} \\ \mathbf{v_2} \\ \vdots \\ \mathbf{v}_m \end{bmatrix} \tag{2.9}$$

If all the PCs are used to represent the original variables, the original raw data matrix is reproduced back as shown in equation (2.9). However, in this situation, the purpose for using Principal Components Analysis as data dimension reduction technique will be lost. In order to maintain the uniqueness of this technique, only

several PCs will be used to represent most of the original data variation. Therefore, if '*a'* of Principal Components are decided to be retained with $a < m$, then equation (2.9) can be adjusted and written as follows:

$$\mathbf{X}_{n\times m} = \mathbf{P}_a\mathbf{V}_a^T + \mathbf{P}_{m-a}\mathbf{V}_{m-a}^T \tag{2.10}$$

$$\mathbf{X}_{n\times m} = \mathbf{P}_a\mathbf{V}_a^T + \mathbf{E} \tag{2.11}$$

The retained principal components [p_1, p_2, ..., p_a], which form the $P_aV_a^T$ term, are associated with systematic variation in data while the residual principal components [p_{a+1}, p_{a+2}, ..., p_m], which form the residual matrix E are considered of containing measurement errors (Seborg *et al.*, 1996). Therefore, PCA is a multivariate analysis technique that could use less number of newly formed variables to represent the original data variations without losing significant information, in which, information here is referred to data variation.

For the graphical representation of PCA, the linear combinations of the original variables in forming the new variables are actually representing selection of a new coordinate system with [P_1, P_2, ..., P_m] as the new axes obtained by rotating the original system with x_1, x_2, ..., x_n as the coordinate axes. The new axes represent the direction with maximum variability and provide a simpler and more parsimonious description of the variance-covariance matrix or correlation matrix (Johnson and Wichern, 1992). Figure 2.1 and 2.2 are prepared to give graphical representations of PCA.

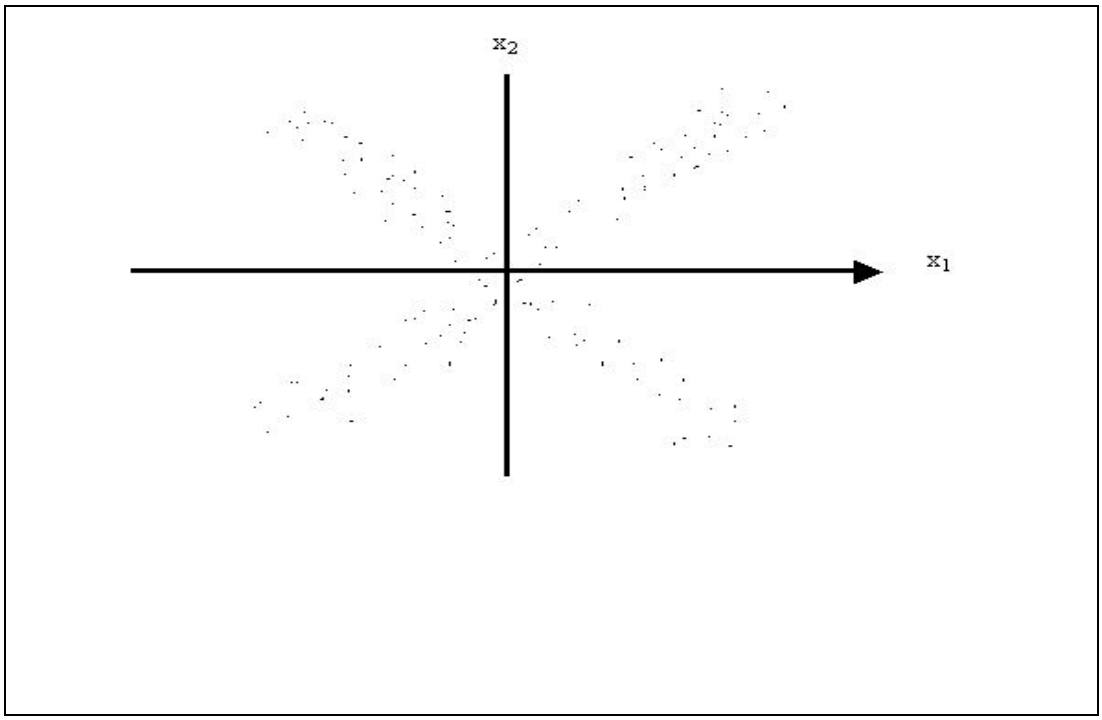

Figure 2.1: System for Two Variables Distribution

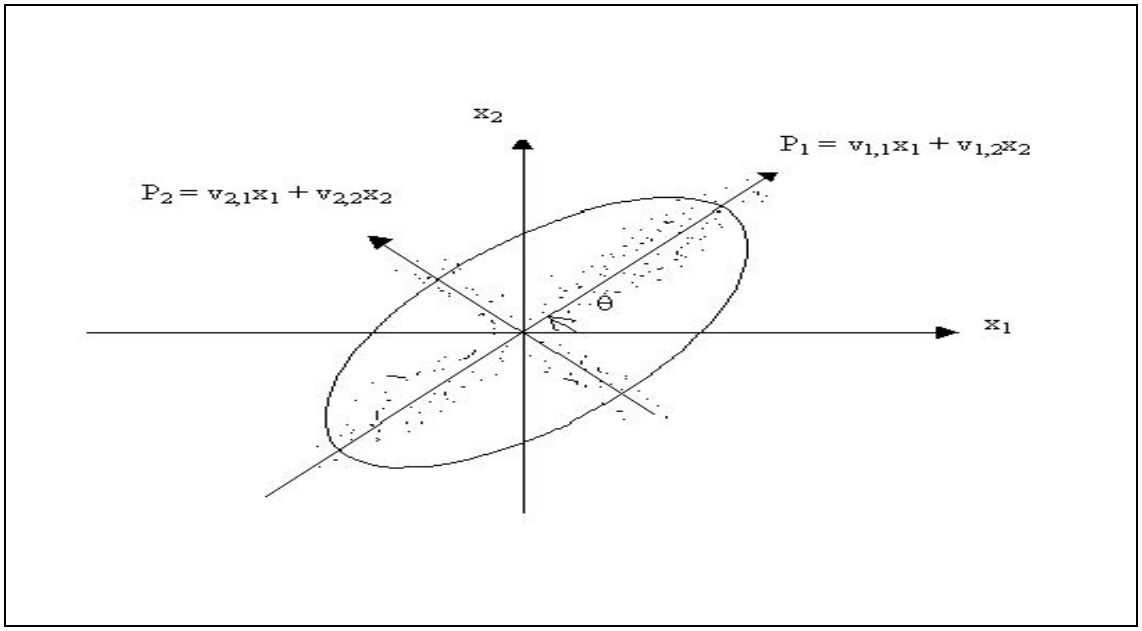

Figure 2.2: Graphical Representation of PCA

2.3 Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is a method used to model the linear relationship between a dependent variable and one or more independent variables. Based on Wise and Gallagher (1996), Linear Regression (MLR) or Ordinary Least Squares (OLS) can be useful for predicting properties of a system based on variables which are only indirectly related to the property. The dependent variable is sometimes also called the predictand, and the independent variables the predictors. MLR is based on least squares where the model is fit such that the sum-of-squares of differences of observed and predicted values is minimized. MLR is probably the most widely used method in dendroclimatology for developing models to reconstruct climate variables from tree-ring series. MLR also suitable to find a single factor that best correlates predictor variables with predicted variables such as concentrations or level. Yuqin et al. (2010) and Pires et al. (2007) state that the purpose of MLR is to model the relationship between two or more explanatory variables and a response variables by fitting a linear equation to observed data. The relationship between target variables, y and explanatory variables, x is modeled by equation below (Green & Carroll, 1976):

$$y_i = b_o + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_k x_{i,k} + e_i \tag{2.11}$$

$where,$

$b_o = regression\ constant$

$K = total\ number\ predictor$

$e_i = error\ term$

$b_k = ceoffocient\ on\ the\ k^{th}\ predictor$

The advantages of MLR are able to analyses attenuated data are more appropriate than correlation analyses because regression coefficients are less influenced by restriction of range of the dependent variables (Cohen, 1975, and Morton, 1982). To perform this method some an assumption must be made in order to obtained an accurate data analysis. Based on Osborne, Jason and Waters (2002), the assumptions are be made such as variables should normally distributed, a linear relationship between the independent and dependent variables must exist. Other than those, the variable should

be measure without an error. MLR is not strictly a "time series" method. The most important point in application to time series is that observations are typically not independent of one another. As a consequence, special attention must be paid to the regression assumption about the independence of the residuals.

2.4 Assumption

The MLR model is based on several assumptions. Provided the assumptions are satisfied, the regression estimators are optimal in the sense that they are unbiased, efficient, and consistent. Unbiased means that the expected value of the estimator is equal to the true value of the parameter. Efficient means that the estimator has a smaller variance than any other estimator. Consistent means that the bias and variance of the estimator approach zero as the sample size approaches infinity. Ostrom (1990, p. 14) lists six basic assumptions for the regression model:

1. **Linearity**: the relationship between the predictand and the predictors is linear. The MLR model applies to linear relationships. If relationships are nonlinear, there are two adjustment should be perform either transform the data to make the relationships linear, or use an alternative statistical model such as binary classification trees. Scatterplots should be checked as an exploratory step in regression to identify possible departures from linearity.
2. **Nonstochastic X**: $E(e_i X_{i,k}) = 0$. The errors are uncorrelated with the individual predictors. This assumption is checked in residuals analysis with scatterplots of the residuals against individual predictors. Violation of the assumption might suggest a transformation of the predictors.
3. **Zero mean:** $E[e_i] = 0$. The expected value of the residuals is zero. This assumption cannot be checked because we have access to the estimated regression residuals, but not to the true unknown errors. The least-squares method used to estimate the regression equation guarantees that the mean of the estimated residuals is zero.
4. **Constant variance:** $E[e_i^2] = \sigma^2$. The variance of the residuals is constant. In time series applications, a violation of this assumption is indicated by some organized pattern of dependence of the residuals on time. An example of violation is a pattern of residuals whose scatter (variance) increases over time. Another aspect of this assumption is that the error variance should not change

systematically with the size of the predicted values. For example, the variance of errors should not be greater when the predicted value of the predictand is large than when the predicted value is small.

5. **Nonautoregression:** $E[e_i e_{i-m}] = 0, m \neq 0$. The residuals are random, or uncorrelated in time. This assumption is one most likely to be violated in time series applications. Several methods of checking the assumption are covered later.
6. **Normality:** The error term is normally distributed. This assumption must be satisfied for conventional tests of significance of coefficients and other statistics of the regression equation to be valid. It is also possible to make no explicit assumption about the form of the distribution and to appeal instead to the Central Limit Theorem to justify the use of such tests. The normality assumption is the least crucial of the regression assumptions.

2.5 Extension of Principal Component Analysis

2.5.1 Multi-Way PCA

Conventional PCA is best for analysing a two-dimensional matrix of data collected from a steady state process, containing linear relationships between the variables. Since these conditions are often not satisfied in practice, several extensions of PCA have been developed. Nomikos and MacGregor (1994) proposed Multi-Way PCA, which allows the analysis of a multi-dimensional matrix. Multi way method organized the data into time ordered block which each represent a single sample or process run. Three dimensional array data (I, batch samples x J, process variables x K, time) is decomposed to two dimensional array (I x JK) data for easier analysis (Wise and Gallagher, 1996). Projection of these three dimensions data into two dimensions makes this method suitable and widely applied for batch processes. Nomikos and MacGregor (1994) used simulated data obtained from a semibatch reactor to monitor the process.

Multi-Way PCA was applied to industrial batch polymerization reactor using Hotelling's T^2 chart for fault detection and contribution plots for fault diagnosis (MacGregor and Kourti, 1995; Nomikos, 1996; Kourti et al., 1996). Multi-Way PCA was also applied using process data collected from an industrial fed-batch fermentation

process (Lennox et al., 2001). Lopes and Monezes (1998) applied this method for an industrial antibiotic production processes to detect faulty batches. Martin et al. (1999) used batch polymerization reactor to illustrate the implementation of Multi-Way PCA. Instead of using T^2 statistic, M^2statistics was used to determine the confidence bound for data not normally distributed. Martin and Morris (1996) proposed M^2 statistics that an empirical density based approach which the bounds calculated is based on density estimation.

Under the Multi-Way PCA monitoring framework, Q statistic is used as the fault detection tool to detect abnormal batch variation whereas contribution plots are used as the fault diagnosis tool to isolate the faulty process variables that are responsible for the out-of-control situation. The disadvantage of contribution plot is that, they cannot isolate the causes automatically without the presence of confidence limit. The plant personnel need to decide whether the out-of-control situation is due to single or multiple faulty process variables.

2.5.2 Multi-Block PCA

Extensions of basic PCA to handle very large processes via Multi-Block PCA were made by MacGregor et al. (1994). This method permits easier modeling and interpretation of a large matrix by decomposing it into smaller matrices or blocks. The Multi-Block PCA enables plant wide monitoring. The monitoring framework for fault detection and diagnosis can be viewed in Figure 2.1.

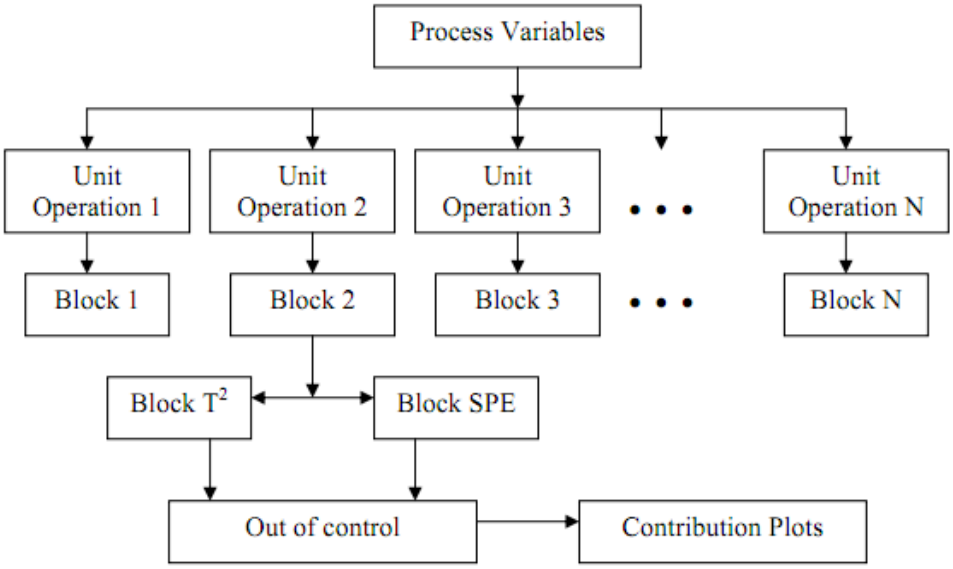

Figure 2.3: Multi-block PCA monitoring framework

The process variables are divided into several blocks with respective to specific unit operation. Block Hotellings' T^2 control chart and Q statistic control chart are used to detect out-of-control situation while contribution plots are used to isolate faulty process variables that cause the out-of-control situation. This approach relies on contribution plots for fault diagnosis; the shortcomings of this method still exist and could not offer complete fault isolation.

2.5.3 Moving PCA

The concept of Moving Principal Components Analysis is based on the idea that a change of correlation in between process variables can be detected by monitoring the directions of principal components (Kano et al., 2000a; Kano et al., 2001b). In order to evaluate the change of direction of each principal component, an index based of the inner product between two principal components is defined. The index proposed in MPCA contained information on the current PCs directions with reference PCs directions to detect any non-conformance situation to the reference models. Previous known fault data sets are used to construct PCA models corresponded to various kinds of fault situation to diagnose the fault cause. The drawbacks of this method are as follow:

i. There are a lot of PCA models, which representing various past fault situations are needed.
ii. Sufficient data have to be collected in order to construct PC models for each fault situation.

2.5.4 Dissimilarity, DISSIM

Kano et al. (2000) proposed monitoring method based on process data distribution known as DISSIM. DISSIM method is based on the idea that a change of operating condition can be detected by monitoring a distribution of time-series data, which reflects the corresponding operating condition. The degree of dissimilarity between data sets is determined in DISSIM method (Kano et al., 2000). Dissimilarity index was defined to evaluate the difference between two data quantitatively. Dissimilarity index control chart is used for fault detection. This index contains the information of the current data distribution with reference data distribution.

For fault diagnosis, each historical known fault data set is used to construct the known fault PCA models, which represents specific known fault data distribution. Besides, a similarity index is introduced to compare the current fault data distribution to each previous known fault PCA models. The proposed method is limited to PCs, which have similar variances because the index cannot function well if the PCs are changed. Other drawbacks of this method are sufficient data is required to construct every known fault PCA models and non-ability to isolate new fault. For fault diagnosis purpose, a contribution of each process variable to the dissimilarity index is introduced for identifying the variables that contribute significantly to an out-of-control value of the index (Kano et al., 2000c). However, there is difficult to identify exact fault causes for the process, which has many feedback control loops and the process variables are complicatedly related to each other.

2.5.5 Multi-Scale PCA

Bakshi (1998) developed Multi-Scale Principal Components Analysis, MSPCA by combining PCA and wavelet analysis. PCA has the ability to extract the relationship between the process variables and de-correlate cross correlation while wavelet analysis has the ability to extract events at different scales, compress deterministic features in a small number of relatively large coefficients, and approximately decorrelate a variety of

stochastic processes (Bakshi, 1999). MS-PCA methodology determines separate PCA models at each scale to identify the scales where significant events occur. MS-PCA method has been applied for fault detection in industrial Fluidized Catalytic Cracker Unit, FCCU. Results showed that MS-PCA detects the faulty condition faster than conventional PCA using Hotelling's T^2statistic and Q statistic but the weakness of this method is that he didn't propose fault diagnosis method.

Kano et al. (2000) applied MS-PCA to monitor problems of a simple two dimension matrix array data obtained from Tennessee Eastman Challenge process. Other researchers, Misra et al. (2002) proposed the combination of PCA and wavelet analysis. In essence, the MS-PCA approach is the same as proposed by Bakshi (1998). However, some differences have been introduced in their study such as multi-scale fault identification technique to identify the type of fault and sensor validation approach to serve as an early warning in case a fault of large magnitude is present. An industrial gas phase tubular reactor system used in this work for process fault diagnosis and sensor fault detection. The outcomes showed that the proposed method was able to detect and identify faults and abnormal events earlier than the conventional PCA approach. The disadvantage of this method is that, it requires basic understanding of the physical and chemical principles governing the process operation to help in clustering the highly correlated variables together before constructing the PCA model. Multi-scale fault identification does not provide the limits for contribution plot.

2.6 Other Applications

The profile monitoring framework includes applications in which numerous measurements of the same variable such as thickness and conversion. Staudhammer, Maness, and Kozak (2007) proposed control charting methods for lumber manufacturing. In their application a laser was used to make dimensional measurements along each board at several locations (Staudhammer, Maness, and Kozak, 2007). It was important to detect process faults which could result in boards not being the desired shape. In this application the key out-of-control shapes in the boards could be specified in advance. Schajer *et al.* (2004) described the measurement system used to obtain the lumber profile data.

Colosimo *et al.* (2007) proposed methods for monitoring dimensional requirements on manufactured items, with a focus on the monitoring of roundness. They considered the monitoring of more general shapes and surfaces. Colosimo *et al.* (2007) compared a principal component approach to a method more commonly used in industry for profile monitoring described by Boeing (1998). The monitoring of shapes is a very promising area of research since the shape of manufactured items is very often an important aspect of quality. These researchers have shown that the standard methods for monitoring shape data are not as efficient as profile monitoring approaches.

Finally, it was pointed out by Woodall (2006) that profile monitoring ideas could be applied in public health surveillance where the interest is in detecting clusters of increased disease rates over time.

3 METHODOLOGY

3.1 Overview

In this chapter the application of MSPC and MLR will be explained step by step on how they being implement which are consisting of two phases.

3.2 Introduction

The complete procedures of fault detection and identification comprise of two main phases namely as off-line modelling and monitoring (phase I) and on-line monitoring (phase II).

The main steps of MSPM system:

i. Fault detection: to designate the departure of observed samples from an acceptable range using a set of parameters.
ii. Fault identification: identifying the observed process variables that are most relevant to the fault which is typically identified by using the contribution plot technique.
iii. Fault diagnosis: specifically determines the type of fault which has been significantly (and should be also validated) contribute to the signal.
iv. Process recovery: remove the cause(s) that contribute to the detected fault.

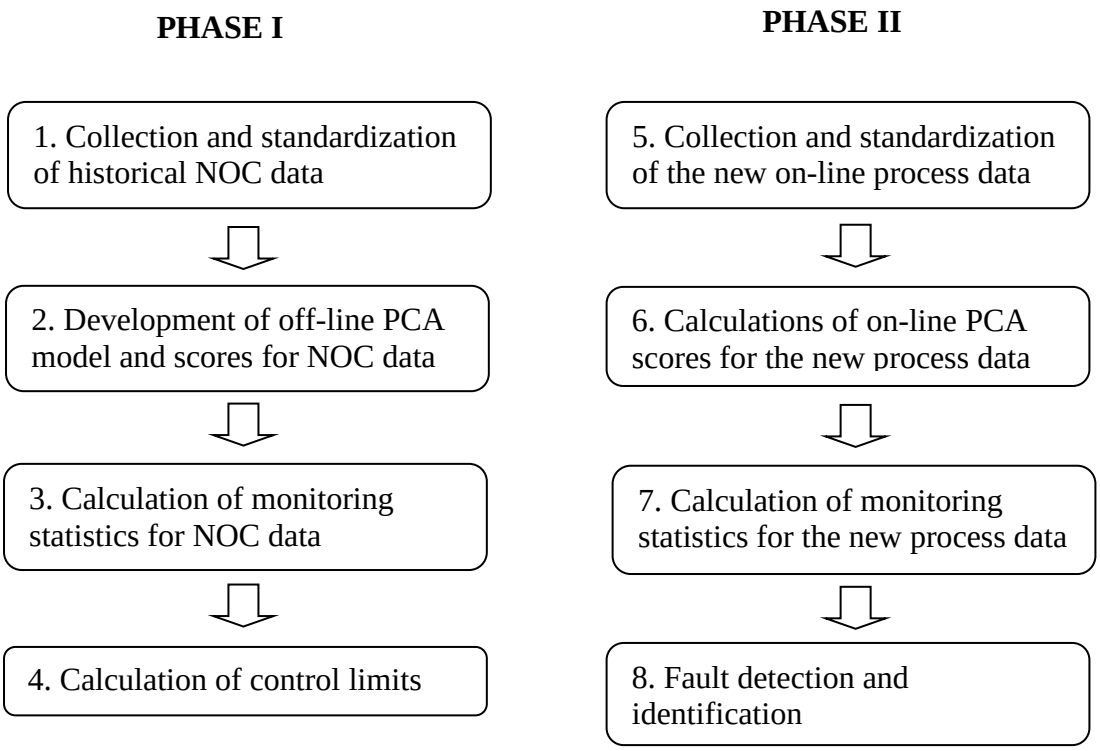

Figure 3.1: Two main phases namely as off-line modelling and monitoring (phase I) and on-line monitoring (phase II)

3.3 Phase I: Off-line Modelling and Monitoring

i. Firstly, a set of normal operation condition (NOC) data, $\mathbf{X}_{n\times m}$ (n: samples, m: variables), are identified off-line based on the historical process data archive.

$$\mathrm{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \tag{3.1}$$

ii. NOC simply implies that the process is operated at the desired setting condition and produces satisfactory products that meet the qualitative as well as quantitative specified standard (Martin et al., 1996).

iii. Then, the data are then standardized to zero mean and unit variance with respective to each of the variables because PCA results depend on data scales.

$$\breve{x}_{j,i} = \frac{\left(x_{j,i} - \bar{x}_i\right)}{\sigma_i} \tag{3.2}$$

iv. In the second step, the development of PCA model for the NOC data requires the establishment of a set of variance-covariance matrix, $\mathbf{C}_{mxm}$.

$$\mathrm{C} = \frac{1}{n-1}\breve{\mathrm{X}}'\breve{\mathrm{X}} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,m} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,m} \end{bmatrix} \tag{3.3}$$

v. **C** is then transformed into a set of basic structures of eigen-based formula.

$$\mathrm{C} = \mathrm{V}\Lambda\mathrm{V}^{\mathrm{T}} \tag{3.4}$$

iiv. Finally, the PCA model of can be simply developed by:

$$\mathbf{P} = \breve{\mathbf{X}}\mathbf{V} \tag{3.5}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{p_1} & \cdots & \mathbf{p_m} \end{bmatrix} = \begin{bmatrix} \breve{x}_{1,1}v_{1,1} + \cdots + \breve{x}_{1,m}v_{m,1} & \cdots & \breve{x}_{1,1}v_{1,m} + \cdots + \breve{x}_{1,m}v_{m,m} \\ \vdots & \cdots & \vdots \\ \breve{x}_{n,1}v_{1,1} + \cdots + \breve{x}_{n,m}v_{m,1} & \cdots & \breve{x}_{n,1}v_{1,m} + \cdots + \breve{x}_{n,m}v_{m,m} \end{bmatrix} \tag{3.6}$$

iiiv. The following equation presents a measure of data variations captured by the first *a* principal components (Jolliffe, 2002).

$$k = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_a}{\lambda_1 + \lambda_2 + \ldots + \lambda_a + \ldots + \lambda_m} \tag{3.7}$$